A Symbolic Dynamics Approach to Volatility Prediction

Peter Tiňo* **Christian Schittenkopf** **Georg Dorffner†**
Austrian Research Institute for Artificial Intelligence
Schottengasse 3, A-1010 Vienna, Austria

Engelbert J. Dockner
Dept. of Business Administration, University of Vienna
Brünner Strasse 72, A-1210 Vienna, Austria

Abstract

We consider the problem of predicting the direction of daily volatility changes in the Dow Jones Industrial Average (DJIA). This is accomplished by quantizing a series of historic volatility changes into a symbolic stream over 2 or 4 symbols. We compare predictive performance of the classical fixed-order Markov models with that of a novel approach to variable memory length prediction (called prediction fractal machine, or PFM) which is able to select very specific deep prediction contexts (whenever there is a sufficient support for such contexts in the training data). We learn that daily volatility changes of the DJIA only exhibit rather shallow finite memory structure. On the other hand, a careful selection of quantization cut values can strongly enhance predictive power of symbolic schemes. Results on 12 non-overlapping epochs of the DJIA strongly suggest that PFMs can outperform both traditional Markov models and (continuous-valued) GARCH models in the task of predicting volatility one time-step ahead.

1 Introduction

Traditionally, option price forecasts are based on implied volatilities derived from an observed series of option prices. The basic assumption behind this approach is that in efficient capital markets with constant volatility of asset returns, the volatility must be reflected in option prices. Taking a different route, Noh, Engle and Kane [14] used a GARCH model [3] to predict the volatility of the rate of return of an asset and then based their predictions of option prices on the GARCH-predicted volatilities. In addition, the volatility change forecasts (volatility is going to increase or decrease) based on historical returns can be interpreted as a buying or selling signal for a straddle. This enables one to implement simple trading strategies to test the efficiency of option markets (e.g. S&P 500 index [14], or German Bund Future Options [5]). If the volatility decreases we go short (straddle is sold), if it increases we take a long position (straddle is bought). In this respect, the quality of a volatility model can be measured by the percentage of correctly predicted directions of volatility change from this period to the next.

In our previous work on predicting the daily volatility of the Austrian stock market index ATX [19], real-valued volatility models were evaluated by considering the squared daily returns the "true" volatilities and comparing them with the model forecasts. In this paper, we take a symbolic dynamics route. The time series of historic volatility changes is quantized into a symbolic sequence characterizing

*also with *Department of Computer Science and Engineering, Slovak University of Technology, Ilkovicova 3, 812 19 Bratislava, Slovakia*

†also with *Dept. of Medical Cybernetics and Artificial Intelligence, Univ. of Vienna*

the original real-valued sequence only through a few distinct events (symbols) such as *sharp increase*, *small decrease*, etc... Instead of modeling the original real valued trajectory, we look for a set of grammatical and probabilistic constraints characterizing its symbolic counterpart.

There have not been many applications of symbolic methods to modeling financial time series. Papageorgiou built predictive models to determine the direction of change in high frequency Swiss franc/U.S. dollar exchange rate (XR) tick data [15] and studied the correlational structure of coupled time series of daily XRs for five major currencies measured against the U.S. dollar [16]. In both cases the real-valued XR returns were quantized into 9 symbols. Papageorgiou predicts the directions of changes in Swiss franc/U.S. dollar XRs using a second order Markov model (MM) and analyses the correlational structure in the five major XRs through a mixed memory MM [18]. Giles, Lawrence and Tsoi [6] considered the same set of five major XRs and predicted the XR directional changes by applying recurrent neural networks to symbolic streams obtained by quantizing the historic real-valued directional change values using the self-organizing map [11].

Generally, it was found that discretization of financial time series can potentially effectively filter the data and reduce the noise. Even more importantly, the symbolic nature of the pre-processed data enables one to interpret the predictive models as lists of clear (and often intuitively appealing) rules. Yet, there are serious shortcomings in using such techniques:

- The determination of the number of quantization intervals (symbols) and their cut values is ad hoc. No strongly supported explanation is given in [15, 16] why 9 symbols with their particular quantization intervals were used. The authors of [6] use up to eight symbols with the cut values determined by the self-organizing map without any attempt to set the quantization intervals to an "optimal"[1] configuration. Kohavi and Sahami [10] warn that naive discretization of continuous data can be potentially disastrous as critical information may be lost due to the formation of inappropriate quantization boundaries. Indeed, discretization should be viewed as a form of knowledge discovery revealing the critical values in the continuous domain.
- Due to (non)stationarity issues, it is a common practice to slide a window containing both training and test sets through the available data thus substantially reducing the amount of training data for model fitting. In such situations using many symbols can be potentionally hazardous as the subsequence statistics is poorly determined. In fact, the results in [6] indicate that the predictive model achieved the best performance with binary input streams.
- Due to training sequence length constraints and the size of used alphabets, the order of MMs is usually set to 2 or 3, whereas only a small set of deeper prediction contexts may really be needed to achieve a satisfactory performance. In case of recurrent neural networks one runs across the well-known vanishing gradient effect [2] reducing the network memory capacity.

We address these issues by

- transforming real-valued time series into symbolic sequences over 2 or 4 symbols. Quantization into 4 symbols is done in an intuitively appealing parametric way.
- using variable memory length Markov models (VLMM) [17] instead of classical fixed-order MMs. VLMMs deal with the familiar explosive increase in the number of MM free parameters (when increasing the model order) by including predictive contexts of variable length with a deep memory *just where it is really needed.*

Apte and Hong [1] address the issue of optimal alphabet size and cut values. They applied a minimal rule generation system R-MINI to monthly S&P 500 data quantized by a special feature

[1]with respect to the performance measure

discretization subsystem. However, the features were quantized prior to the rule generation process without any reference to the final model's predictive behavior. The R-MINI rules are in disjunctive normal form.

Bühlmann [4] models the extreme events of return of the Dow Jones and volume of the NYSE given their previous histories. The original return and volume series are quantized into streams over 3 ordinal categories (*lower extreme, usual, upper extreme*) that are used to fit a hierarchy of generalized linear models viewed as sieve approximations to a finite state Markov chain. The cut values for the 3 categories correspond to the 2.5% and 97.5% sample quantiles, so that the *lower* and *upper extreme* categories describe extreme events with expected overall occurrence of about 5%. We adopt this quantization strategy (in a slightly modified form) in our experiments.

.

2 Prediction system

The main idea in building a variable memory length (VLMM) Markov model on a given input stream S over an alphabet $\mathcal{A} = \{1, 2, ..., A\}$ is to consider all predictive contexts[2] $w \in \mathcal{A}^l$, $1 \leq l \leq L$, up to a certain pre-specified depth L and select only contexts v the predictive (empirical) distributions $P(s|v)$, $s \in \mathcal{A}$, of which differ significantly from those of their suffices, e.g. include an extended context au, $a \in \mathcal{A}$, $u \in \mathcal{A}^l$, $1 \leq l \leq L-1$, if the Kullback-Leibler divergence between the next-symbol distributions for the candidate prediction contexts u and au, weighted by the prior (empirical) distribution of the extended context au, exceeds a given threshold ϵ [17]

$$P(au) \sum_{s \in \mathcal{A}} P(s|au) \log \frac{P(s|au)}{P(s|u)} \geq \epsilon. \tag{1}$$

The parameters ϵ and L are supplied by the modeler. We have shown [13] that constructing VLMMs in this manner can be very troublesome. The construction parameters determine the VLMM size only implicitly without any intuitive specific relation and often one has to spend a fair amount of time when constructing a series of VLMMs of increasing size.

In an attempt to deal with this problem we introduced a novel class of predictive models, called prediction fractal machines (PFMs) [13], similar in spirit to VLMMs, the construction of which is fast and intuitive in the model size. Construction of PFMs starts by transforming the n-block structure of the training sequence $S = s_1 s_2 ...$ over $\mathcal{A}$ into a spatial structure of points in a unit hypercube, called the chaos n-block representation of S. The n-blocks $u = u_1 u_2 ... u_n \in \mathcal{A}^n$ are represented as points

$$u(x) = u_n(u_{n-1}(...(u_2(u_1(x)))...)) = (u_n \circ u_{n-1} \circ ... \circ u_2 \circ u_1)(x), \quad x \in X, \tag{2}$$

where $X = [0, 1]^N$, $N = \lceil \log_2 A \rceil$, and the maps $1, 2, ..., A$,

$$i(x) = kx + (1-k)t_i, \quad t_i \in \{0, 1\}^N, \quad t_i \neq t_j \text{ for } i \neq j, \tag{3}$$

act on X with a contraction coefficient $k \in (0, \frac{1}{2}]$. For $Y \subseteq X$, $u(Y) = \{u(x) | \; x \in Y\}$.

Denote the center $\{\frac{1}{2}\}^N$ of X by x_*. The chaos n-block representation of S is a sequence of points

$$CBR_{n,k}(S) = \left\{ S_i^{i+n-1}(x_*) \right\}_{i \geq 1}, \tag{4}$$

containing a point $w(x_*)$ for each n-block $S_i^{i+n-1} = s_i s_{i+1} ... s_{i+n-1} = w$ in S. The map $w \rightarrow w(x_*)$ is one-to-one.

[2] To keep the construction procedure feasible it is common to consider only contexts with probability of occurrence greater than some pre-determined threshold θ.

The chaos n-block representation of symbolic sequences is related to the chaos game representation of DNA sequences introduced by Jeffrey [9] and has many useful properties. We proved [12] that the estimates of generalized Rényi dimension spectra, quantifying the multifractal scaling properties of $CBR_{n,k}(S)$, directly correspond to the estimates of the Rényi entropy rate spectra measuring the statistical structure in the sequence S. In particular, for infinite sequences S, as the block length n grows, the box-counting fractal dimension and the information dimension estimates on $CBR_{n,k}(S)$, tend to the sequences' topological and metric entropies, respectively, scaled by $(\log \frac{1}{k})^{-1}$.

The chaos n-block representation codes the n-block suffix structure in the following sense [12]: if $v \in \mathcal{A}^+$ is a suffix of length $|v|$ of a string $u = rv$, $r, u \in \mathcal{A}^+$, then $u(X) \subset v(X)$, where $v(X)$ is an N-dimensional hypercube of side length $k^{|v|}$. Hence, the longer is the common suffix shared by two n-blocks, the closer the n-blocks are mapped in the chaos n-block representation $CBR_{n,k}(S)$. On the other hand, the Euclidean distance between points representing two n-blocks u, v, that have the same prefix of length $n-1$ and differ in the last symbol, is at least $1-k$.

Suppose our model cannot have more than M prediction contexts. A natural smoothness constraint that L-blocks with long common suffices are likely to produce similar continuations, whereas L-blocks with different suffices may lead to different future scenarios leads us to the idea that L-blocks (potential predictive contexts of maximal depth) should factorize the set $[S]_L$ of allowed L-blocks in the training sequence S into a set of M equivalence classes, such that blocks within each equivalence class share as long a common suffix as possible. Recalling the suffix structure coding properties of the chaos L-block representation $CBR_{L,k}(S)$ of the training sequence S, this corresponds to partitioning the $CBR_{L,k}(S)$ into M subsets, each of diameter as small as possible. In practical terms, this means allocation of points from $CBR_{L,k}(S)$ to M codebook vectors $b_1, ..., b_M \in X$, such that the loss

$$E(S) = \sum_{w \in [S]_L} P(w)\ d_E^2(w(x_*), c(w)) \tag{5}$$

is minimal, where $c(w) \in \{b_1, ..., b_M\}$ is the codebook vector to which the point $w(x_*)$ is allocated, and d_E is the Euclidean distance.

The prediction probabilities in PFMs are determined by

$$\mathcal{P}(s|b_i) = \frac{N(i,s)}{\sum_{a \in \mathcal{A}} N(i,a)}, \quad s \in \mathcal{A}, \tag{6}$$

where $N(i,a)$ is the number of $(L+1)$-blocks wa, $w \in \mathcal{A}^L$, $a \in \mathcal{A}$, in the training sequence, such that the point $w(x_*)$ is allocated to the codebook vector b_i.

Given a history $w \in \mathcal{A}^L$ of L symbols, the next symbol distribution provided by the PFM is $\mathcal{P}(s|c(w))$, with $c(w)$ defined in (5).

3 Experiments

The time series $\{x_t\}$ of the Dow Jones Industrial Average (DJIA) from Feb. 1 1918 until Dec. 31 1997 (21620 measurements) was transformed into a time series of returns $r_t = \log x_{t+1} - \log x_t$ and divided into 13 epochs (as shown in figure 1) each containing 1663 values (spanning approximately 6 years). The series $\{r_{t+1}^2 - r_t^2\}$ of differences between the successive squared returns is quantized with respect to the direction of daily volatility change into a binary sequence $\{C_t\}$

$$C_t = \begin{cases} 1 \text{ (down)}, & \text{if } r_{t+1}^2 - r_t^2 < 0 \\ 2 \text{ (up)}, & \text{otherwise.} \end{cases} \tag{7}$$

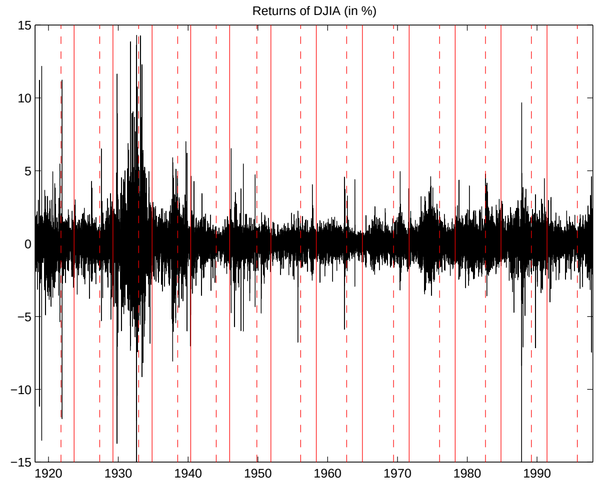

Figure 1: Returns of the DJIA form Feb.1 1918 until Dec. 31 1997. The data is divided into 13 epochs (dotted vertical lines). Each epoch is split into a training sequence and a test sequence (dashed vertical lines).

Since within each epoch, *downs* (1s) and *ups* (2s) occur with approximately the same frequency, sequences $\{D_t\}$ over 4 symbols are obtained by quantizing the series $\{r_{t+1}^2 - r_t^2\}$ as follows:

$$D_t = \begin{cases} 1 \text{ (extreme down)}, & \text{if } r_{t+1}^2 - r_t^2 < \theta_1 < 0 \\ 2 \text{ (normal down)}, & \text{if } \theta_1 \leq r_{t+1}^2 - r_t^2 < 0 \\ 3 \text{ (normal up)}, & \text{if } 0 \leq r_{t+1}^2 - r_t^2 < \theta_2 \\ 4 \text{ (extreme up)}, & \text{if } \theta_2 \leq r_{t+1}^2 - r_t^2, \end{cases} \tag{8}$$

where the parameters θ_1 and θ_2 correspond to $\mathcal{Q}$ percent and $(100 - \mathcal{Q})$ percent sample quantiles, respectively. So, the upper (lower) $\mathcal{Q}$% of all daily volatility increases (decreases) in the sample are considered extremal, and the lower (upper) $(50 - \mathcal{Q})$% of daily volatility increases (decreases) are viewed as normal. In our experiments $\mathcal{Q} \in \{5, 10, 15, ..., 45\}$.

We performed two experiments. The first experiment identifies the amount of detectable memory in the daily volatility change process. In the second experiment we simulate a realistic setting, where one is forced to split the available data into separate training and validation sets, the latter used for selection of the most appropriate model size M and quantization quantile $\mathcal{Q}$ for the models trained on the training set.

In [13] we have shown that our prediction fractal machines (PFMs) and variable memory length Markov models (VLMM) outperform the classical fixed-order Markov models (MMs) on chaotic sequences with deep memory structure. In such cases it is beneficial, instead of simply considering all prediction contexts of certain length, to accept only a limited set of carefully selected long contexts. On the other hand, a shallow memory structure in sequences reduces the advantage of using VLMMs (or PFMs) instead of the classical MMs. In the experiments reported here, the length of the window in constructing the chaos block representations of the training sequences was set to $L = 20$ and the contraction factor is $k = \frac{1}{2}$.

In the first experiment, for each epoch, we split the data into a training sequence (the first 1100 symbols) and a test sequence (the rest of the epoch). The two sequences are quantized according to the extremal event quantile $\mathcal{Q}$ (see (8)). Then, PFMs and MMs of increasing size are trained on the training set and tested on the test set with respect to the percentage of correct guesses of the

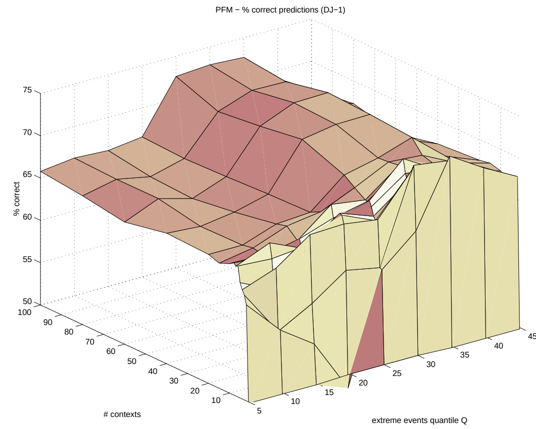

Figure 2: Performance of PFMs on the first DJIA data epoch measured by the percentage of correct volatility direction predictions on out-of-sample data.

volatility change direction for the next day[3]. If the next symbol in the test set is 1 or 2 (3 or 4) and the sum of conditional next symbol probabilities for 1 and 2 (3 and 4) given by a model is greater than 0.5, the model guess is considered correct. Figures 2 and 3 show the performance of PFMs and MMs, respectively, on the out-of-sample data in the first epoch. The 35% quantile $\mathcal{Q}$ for the extremal events seems to work best for both models. PFMs achieve the best performance with very few prediction contexts (up to 10). Further increase in the model size does not give any improvement in the prediction performance. The performance of MMs improves up to a memory of order 3 and then the MMs clearly overfit the training set (the performance on the test set deteriorates). Figures 4 and 5 suggest that similar observations can be made for the remaining 12 epochs. Depending on the epoch, the optimal value for the quantile $\mathcal{Q}$ ranges from 15 to 40, and the best-performing MM order varies from 1 to 3. In general, in accordance with intuition, more volatile epochs need deeper memory and larger extreme events quantile $\mathcal{Q}$.

It is interesting that PFMs do not identify deep dominant prediction contexts that would lead to a highly superior performance (compared to MMs) as is, for example, the case with the chaotic Feigenbaum sequence [13]. This supports the widely held belief [8] that noise is a dominant component in financial data and that a direct use of techniques known from chaotic dynamics[4] does not bring any dramatic improvement in data description and predictability.

In the second experiment we take a more realistic view. If we were to use our models in a day-to-day trading, we would need to decide (based on the historic data) what extreme event quantile $\mathcal{Q}$ and model size to use for future predictions. The experimental setting from the first experiment is slightly modified in that the test sets from the first experiment are now validation sets on which we select, using the models fitted on the training sets[5], the optimal (model size,quantile $\mathcal{Q}$) combination. The selected models are then tested on test sets, which are in this case the first 600 symbols from the next epoch.

The results are summarized in figure 6. We also plot the performance of a simple 2 symbol (first-order) reversal strategy (dotted line) predicting *up* whenever the previous move was *down* and vice-versa. 4 symbol quantization schemes (solid line) yield consistently better performance than their 2 symbol counterparts (dashed line). Moreover, PFMs (squares) working under 4 symbol schemes

[3] Reported results for PFMs are average values across 10 vector quantization runs (K-means clustering) in the PFM construction

[4] based on the assumption that there is no (or very little) noise affecting the dynamics of the system

[5] the training sets remain unchanged

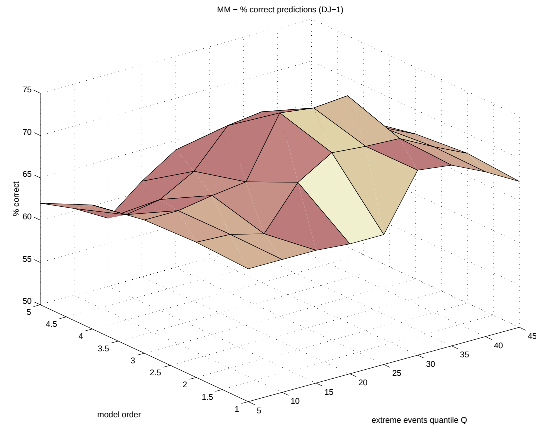

Figure 3: Performance of the classical MMs on the first DJIA data epoch measured by the percentage of correct volatility direction predictions on out-of-sample data.

tend to have better performance than classical MMs (stars). The situation is reversed in the 2 symbol case, where the validation strategy does not prevent PFMs from including a few over-specialized contexts. PFMs rarely perform better than the reversal strategy. This is partially caused by the sample distribution that varies from epoch to epoch. Regime shifts in data distribution can be dangerous for potentially highly specialized models like PFMs or VLMMs. In real trading, the best strategy would be to often re-estimate the predictive model as the new samples arrive.

We also tested the approaches presented in [14, 5] that use GARCH-based methods to model the volatility of the returns and then plug the volatility change forecasts into trading strategies. GARCH [3] and GJR [7] are among the most widely used (real-valued) models of volatility. Basic to these models is the notion that the financial time series $\{x_t\}$ under study can be decomposed into a predictable component μ_t and an unpredictable component e_t, which is assumed to be zero mean Gaussian noise of finite variance σ_t^2: $x_t = \mu_t + e_t$. The models are thus characterized by time-varying conditional variances σ_t^2 and are therefore well suited to explain volatility clusters. The conditional mean is often modelled as a linear function of the previous value: $\mu_t = a x_{t-1}$. For a GARCH(p,q) model the conditional variance σ_t^2 is given by

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i e_{t-i}^2 + \sum_{i=1}^{p} \beta_i \sigma_{t-i}^2. \tag{9}$$

The GJR model is an extension of the GARCH(1,1) model, and it incorporates also asymmetric effects:

$$\sigma_t^2 = \alpha_0 + \alpha_1 e_{t-1}^2 + \alpha_2 \phi_{t-1} e_{t-1}^2 + \beta_1 \sigma_{t-1}^2 \tag{10}$$

where $\phi_{t-1} = 1$ if $e_{t-1} < 0$ and $\phi_{t-1} = 0$ otherwise. GJR models are motivated by the fact that stock returns are characterized by a leverage effect, i.e. volatility increases as returns for stocks decrease. In each epoch, the GARCH(1,1), GARCH(2,2) and GJR models were fit to the training set in the maximum likelihood framework. Following [19], the fitted models were evaluated by considering the squared returns r_t^2 the "true" volatility and comparing them with the forecasted volatility σ_t^2. A prediction is classified correct if $(\sigma_t^2 - r_{t-1}^2)(r_t^2 - r_{t-1}^2) > 0$. As with the symbolic models, based on the validation set performance, we select a GARCH-based candidate and test it on the first 600 points from the following epoch. The results (dashed-dotted line with circles in figure 6) indicate that, compared with symbolic models, the GARCH-based models are more sensitive to shifts in probability regimes of returns. The worst performances of these models (epochs 1, 6 and 12) correspond to cases

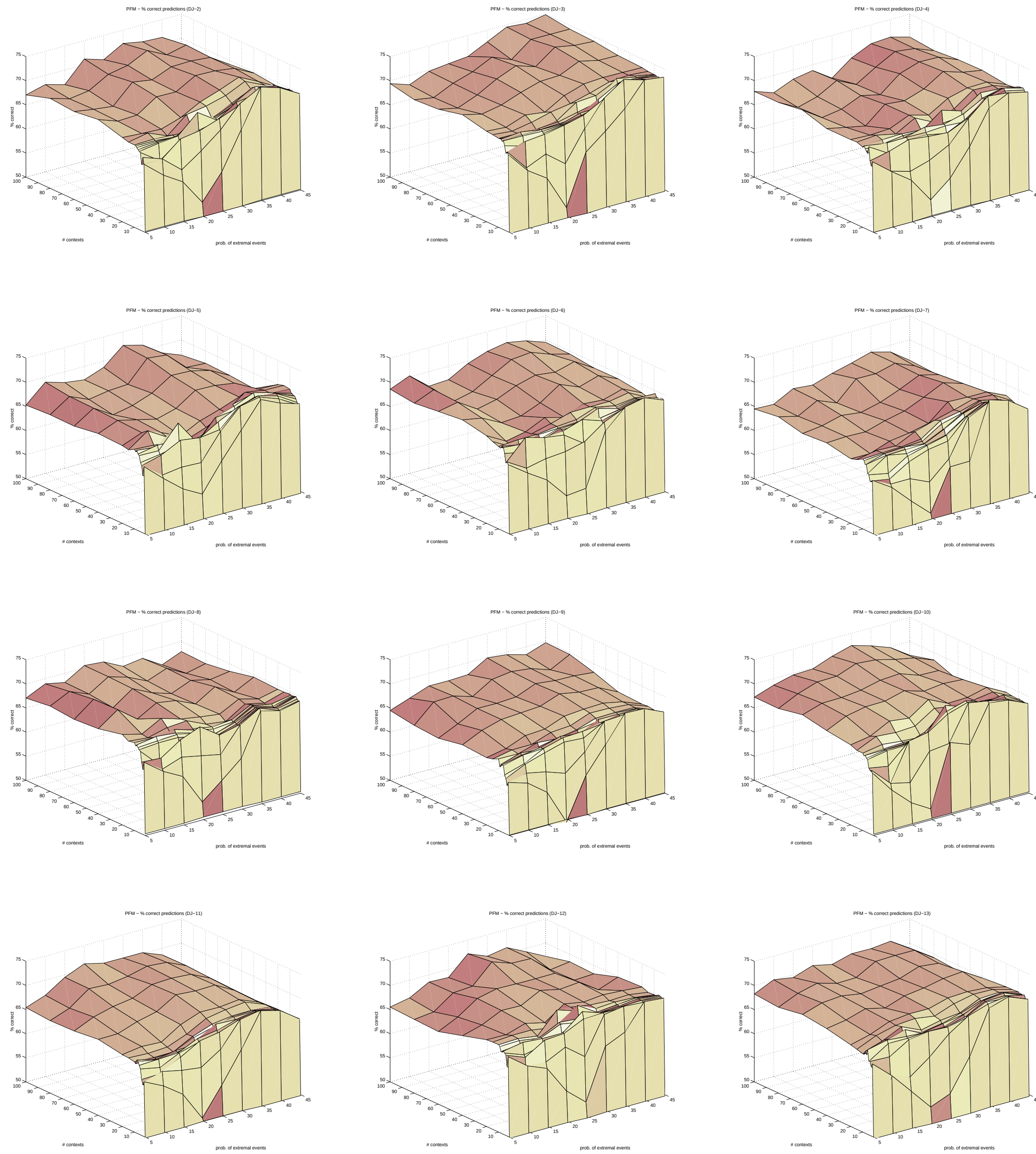

Figure 4: Performance of PFMs on the DJIA data epochs 2–13. (upper-left to bottom-right).

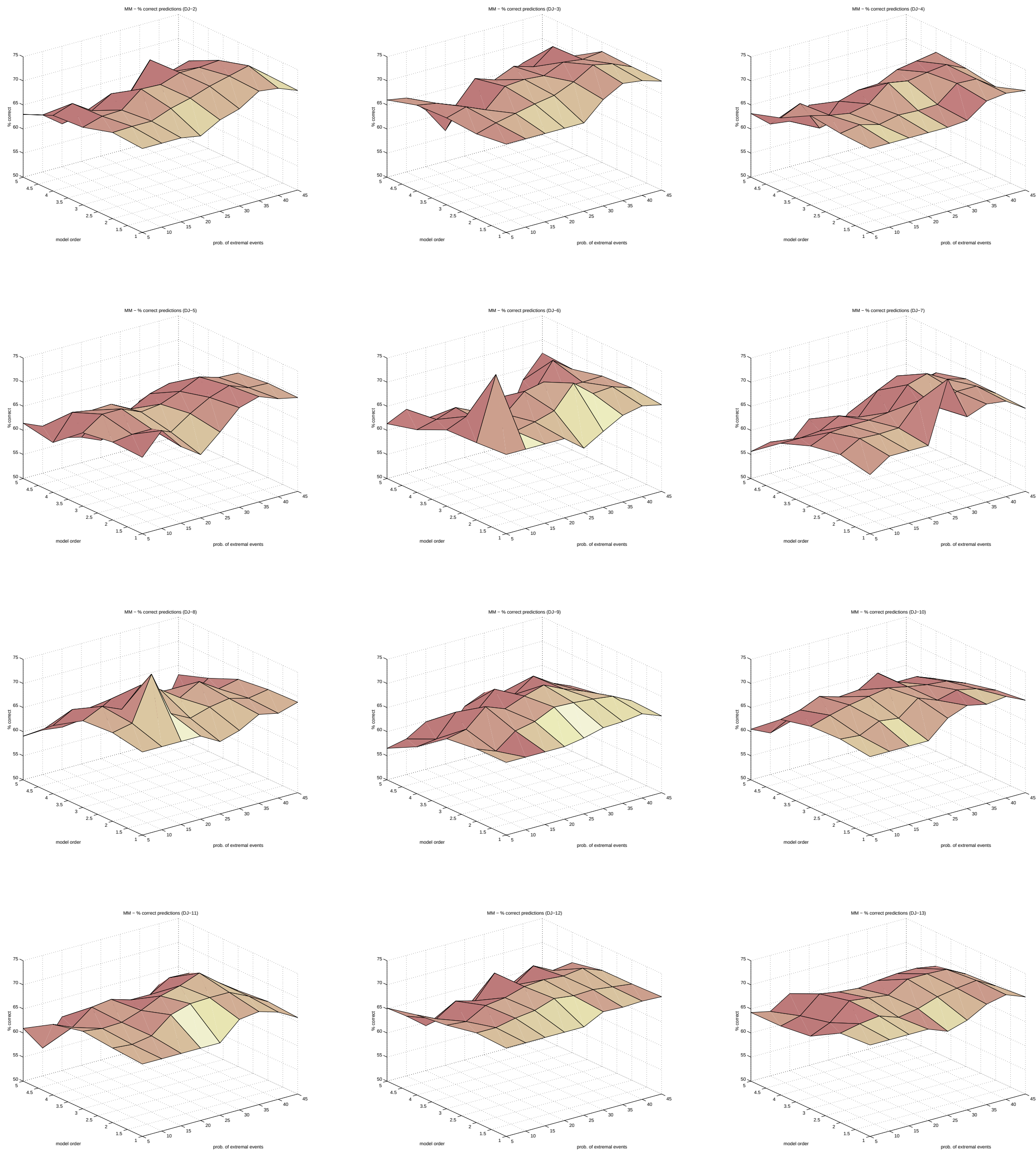

Figure 5: Performance of the classical MMs on the DJIA data epochs 2–13. (upper-left to bottom-right).

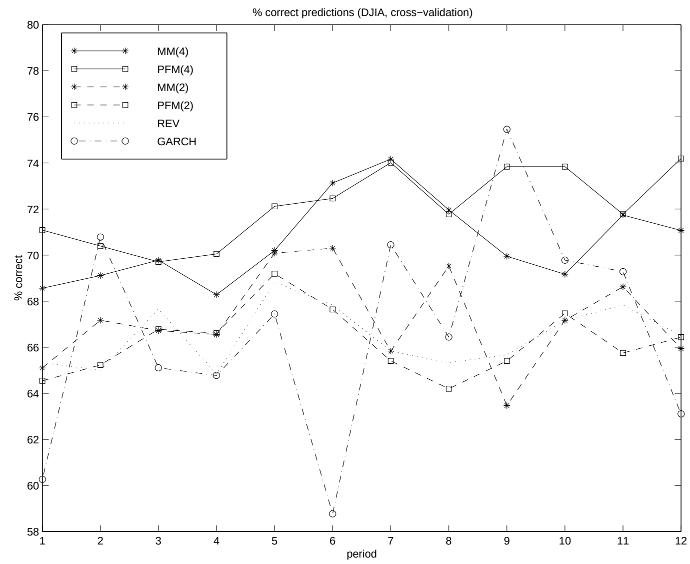

Figure 6: Prediction performance of the PFM (squares) and MM (stars) validation set candidates. The performance is evaluated on the test set containing the first 600 points of the next epoch. Results corresponding to 4 and 2 symbol quantization schemes are shown as solid and dashed lines, respectively. The dotted line corresponds to the simple reversal strategy. Also shown is the performance of the GARCH-based candidates (dashed-dotted line with circles).

where the training set distribution seem to differ from that of the test set (the first 600 points in the next period). We stress, however, that the GARCH-based models were evaluated with respect to volatility change direction predictions, whereas they were trained with a different perspective in mind - to model the conditional distribution of returns, so that the training data are likely to be generated from the model distribution. We included the GARCH experiment for a comparison with the previous approaches to volatility change direction prediction using real-valued models.

In order to test the differences between the three major methods, we performed several significance tests. A t-test over the twelve percentage values indicated that the PFM is significantly better than both the MM ($p < 0.05$) and GARCH ($p < 0.005$). A Wilcoxon matched-pairs signed-ranks test also revealed significant differences between the PFM and MM ($p < 0.05$) and GARCH ($p < 0.005$), respectively. Finally, we performed a McNemar test on the two pairs of classifiers over the twelve independent test sets separately. In eight of the ten test sets where PFM showed better classification performance than the GARCH models, the test indicated significance. For the PFM vs. the MM, this was the case in 4 out of 8 cases. For none of the test sets, either MM or GARCH was ever significantly better than the PFM. One must add, however, that for all three tests, some assumptions appear to be at least mildly violated (non-Gaussian distributions, skewness of differences, possible dependencies between observations).

4 Conclusion

We have investigated potential benefits of using ideas from symbolic dynamics in prediction of daily volatility changes. The two key points of this contribution are

1. the use of a simple parametric scheme to find the best quantization cut values in a data-driven fashion

2. the use of Markov models with variable length contexts to test the amount of detectable memory in the process.

Although the results from the significance tests have to be taken with caution, the results strongly suggest

- that there appears to indeed lie some potential in using a symbolic dynamics approach to volatility forecasting, in which continuous values are replaced by symbols through quantization, and
- that our novel prediction fractal machine can outperform more traditional Markov models in the prediction task.

In addition, the results suggest that daily volatility changes of the DJIA only show rather shallow finite memory structure (both MM and PFM tended to be of depth 2 or 3). 4 symbol quantization schemes, *when the cut values are carefully selected*, yield better performance than the 2 symbol schemes. This (together with the fact that in 4 symbol schemes variable memory length models tend to slightly outperform the classical MMs) suggests that there is a grammatical structure in the data to be grasped, but one cannot use ad hoc quantization techniques.

Of course, implementation and evaluation of different trading strategies using predictions of volatility changes given by various models would be highly desirable (this work has already been started). We did not investigate quantization schemes using more than 4 symbols, since the space of possible quantization parametrizations would rapidly grow. It may be an interesting topic for the future research, though. Also, in this study, we did not discard the rules with little predictive power, i.e. rules that predict *ups* with only slightly higher probability than *downs* and vice-versa. Discarding such rules would lead to no-trading situations as in [15]. We plan to continue our research in this direction. Rules will be discarded according to a special discard parameter. The value of the discard parameter should reflect a balance between the number of rules we are willing to get rid of and the number of trading steps we should perform in order to accumulate a profit. Trading strategies using our predictive models will play a prominent role in this study.

Acknowledgements

This work was supported by the Austrian Science Fund (FWF) within the research project "Adaptive Information Systems and Modeling in Economics and Management Science" (SFB 010). The Austrian Research Institute for Artificial Intelligence is supported by the Austrian Federal Ministry of Science and Transport.

We thank our colleagues Arthur Flexer and Peter Sykacek for providing help with the significance tests. Furthermore we are grateful to Fritz Leisch and Andreas Weingessel for valuable discussions.

References

[1] C. Apte and S.J. Hong. Predicting equity returns from securities data. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 541–560. AAAI/MIT Press, 1994.

[2] Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166, 1994.

[3] T. Bollerslev. A generalized autoregressive conditional heteroscedasticity. *Journal of Econometrics*, 31:307–327, 1986.

[4] P. Buhlmann. Extreme events from return-volume process: adiscretization approach for complexity reduction. *Applied Financial Economics, to appear*, 1998.

[5] E.J. Dockner, G. Strobl, and A. Lessing. Volatility forecasts and the profitability of option trading strategies. Technical report, University of Vienna, Austria, 1988.

[6] C.L. Giles, S. Lawrence, and A.C. Tsoi. Rule inference for financial prediction using recurrent neural networks. In *Proceedings of the conference on Computational Intelligence for Financial Engineering, New York City, NY*, pages 253–259, 1997.

[7] L.R. Glosten, R. Jagannathan, and D.E. Runkle. On the relation between the expected value and the volatility of the nominal excess return on stocks. *Journal of Finance*, 48:1779–1801, 1993.

[8] T. Jaditz and C.L. Sayers. Is chaos generic in economic data? *Int. Journal of Biffurcation and Chaos*, 3(3):745–755, 1993.

[9] J. Jeffrey. Chaos game representation of gene structure. *Nucleic Acids Research*, 18(8):2163–2170, 1990.

[10] R. Kohavi and M. Sahami. Error-based and entropy-based discretization of continuous features. In E. Simondis, J. Han, and U. Fayyad, editors, *Proceedings of the Second International Conference on Knowledge Discovery in Databases*, pages 114–119. AAAI Press, 1996.

[11] T. Kohonen. The self–organizing map. *Proceedings of the IEEE*, 78(9):1464–1479, 1990.

[12] P. Tiňo. Spatial representation of symbolic sequences through iterative function systems. Technical Report TR-98-17, Austrian Research Institute for Artificial Intelligence, Austria, 1998.

[13] P. Tiňo and G. Dorffner. Constructing finite-context sources from fractal representations of symbolic sequences. Technical Report TR-98-18, Austrian Research Institute for Artificial Intelligence, Austria, 1998.

[14] J. Noh, R.F. Engle, and A. Kane. Forecasting volatility and option prices of the s&p 500 index. *Journal of Derivatives*, pages 17–30, 1994.

[15] C.P. Papageorgiou. High frequency time series analysis and prediction using markov models. In *Proceedings of the conference on Computational Intelligence for Financial Engineering, New York City, NY*, pages 182–185, 1997.

[16] C.P. Papageorgiou. Mixed memory markov models for time series analysis. In *Proceedings of the conference on Computational Intelligence for Financial Engineering, New York City, NY*, pages 165–170, 1998.

[17] D. Ron, Y. Singer, and N. Tishby. The power of amnesia. In *Advances in Neural Information Processing Systems 6*, pages 176–183. Morgan Kaufmann, 1994.

[18] L.K. Saul and M.I. Jordan. Mixed memory markov models. In *Proceedings of the 6th International Workshop on Artificial Intelligence and Statistics, Fort Lauderdale, Florida*, 1998.

[19] C. Schittenkopf, G. Dorffner, and E.J. Dockner. Volatility prediction with mixture density networks. In *Proceedings of the International Conference on Artificial Neural Networks, accepted*, Skövde, Sweden, 1998.